Performance of Vocabulary-Independent Speech Recognizers with Speaker Adaptation

*Oh Wook Kwon, **Chong Kwan Un, and *Hoi Rin Kim

Abstract

In this paper, we investigated performance of a vocabulary-independent speech recognizer with speaker adaptation. The vocabulary-independent speech recognizer does not require task-oriented speech databases to estimate HMM parameters, but adapts the parameters recursively by using input speech and recognition results. The recognizer has the advantage that it relieves efforts to record the speech databases and can be easily adapted to a new task and a new speaker with different recognition vocabulary without losing recognition accuracies. Experimental results showed that the vocabulary-independent speech recognizer with supervised offline speaker adaptation reduced 40% of recognition errors when 80 words from the same vocabulary as test data were used as adaptation data. The recognizer with unsupervised online speaker adaptation reduced about 43% of recognition errors. This performance is comparable to that of a speaker-independent speech recognizer trained by a task-oriented speech database.

I. Introduction

Recently, speaker-independent speech recognizers have shown remarkable recognition accuracies by virtue of sophisticated acoustic modeling and large speech databases. However, their performances are yet inferior to speaker-dependent speech recognizers. To recognize speech utterances in dynamically-varying dialogue contexts or situations, a vocabulary-independent (variable-vocabulary) speech recognizer [1] is desirable. In that case, a speech recognizer trained by a task-oriented speech database shows poor performance because new vocabulary different from training vocabulary has to be recognized. On the contrary, in the vocabulary-independent speech recognizer, we can use different recognition vocabulary according to dialogue contexts.

The vocabulary-independent speech recognizer has the advantage that it requires no task-oriented training speech databases and relieves efforts to record to speech databases, so that it can be easily adapted to a new task and a new speaker with different recognition vocabulary without losing recognition accuracies. The recognizer can be used where recognition vocabulary can not be predefined and has to be changed dynamically with varying dialogue contexts as in speech-driven web browsers. The recognizer can be also used in a dictator application, where the recognizer are initially trained by phonetically-balanced speech data and then are adapted to a new speaker. In this paper, we aim to improve the performance of the vocabulary-independent speech recognizer using speaker adaptation techniques to the level comparable to the task-oriented speech recognizer.

Three approaches have been studied to adapt continuous density hidden Markov model (HMM)-based speech recognizers to a new speaker: A maximum *a posteriori* (MAP) estimation approach, a transform approach, and a smoothing approach.

In the MAP estimation approach [2], [3], [4], [5], a parameter is estimated by maximizing the posterior probability of the parameter given observed samples. When the prior density for a parameter is Gaussian, the resultant estimation formula is usually represented by a weighted sum of the prior parameter and the sample parameter computed by maximum likelihood estimation [6]. And the weight is determined according to the number of observed samples and the variance of the prior density of the parameter. When adaptation data (observed samples) are sufficient for all HMM parameters, the speech recognizer adapted by the MAP estimation approach converges to a speaker-dependent speech recognizer, which can be regarded as optimal. However, it is desirable to use as short adaptation words as possible to relieve a user's efforts to pronounce adaptation utterances. Hence, adaptation speech

*Spoken Language Processing Section, Electronics and Telecommunications Research Institute
** Communications Research Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology

Manuscript Received, May 20, 1997.

data are usually insufficient and consequently some parameters are short of observed samples for reliable estimation. To solve this problem, an extended MAP algorithm [5] was proposed, where correlations among parameters were exploited to estimate parameters reliably with small adaptation data.

Hyperparmeters [4] of the prior density used in MAP estimation are often assumed known. But, in reality, they should be estimated also from training data. Therefore, the MAP estimation approach causes a new problem of estimating hyperparameters of the prior density. In this sense, the extended MAP algorithm has the disadvantage that it requires a large number of hyperparameters to be estimated and it also increases computational complexity. In practice, heuristic or approximate methods are often used to estimate the hyperparameters.

In the transform approach, a global transform for each class that maximizes the probability of the observed samples is estimated [7], [8]. Compared with the MAP estimation approach, this approach shows good performance with small adaptation data, but shows poor performance with sufficient observed samples. The performance of the recognizer is severely affected by the number of classes to be determined empirically. The performance does not converge to that of a speaker-dependent speech recognizer with sufficient adaptation data.

In the smoothing approach, an estimated parameter is obtained in two steps. In the first step, difference is calculated between the parameter obtained by maximum likelihood estimation and the corresponding prior parameter. Then, the difference is smoothed by adjacent differences to the parameter by using the concepts of the vector field theory [9] or the Markov random field theory [10]. The smoothing techniques inherently resemble fuzzy smoothing. The degree of smoothing should be carefully determined for proper working. This approach requires a large number of parameters to be determined empirically and therefore increases computational complexity.

In this paper, we adopt the MAP estimation approach because it has a simple structure and enables the recognizer to converge to a speaker-dependent one with sufficient amount of adaptation data. Then we simplify the MAP estimation algorithm to reduce the number of estimated parameters so that it can be easily applied to an existing speech recognizer without making an effort to estimate the corresponding hyperparameters. Experimental results showed that the vocabulary-independent speech recognizer adapted by using the same vocabulary as test data reduced 40% of recognition errors. And the recognizer adapted by using vocabulary different from test data yielded recognition results worse than the recognizer without speaker adaptation. The performance of the vocabulary-independent speech recognizer with unsupervised online speaker adaptation was comparable to that of a speaker-independent speech recognizer trained by a task-oriented speech database.

Following the introduction, a simplified algorithm to adapt HMM parameters is described in Section II. In Section III, a vocabulary-independent speech recognizer is explained. In Section IV, experimental results and discussion are given. Finally, we summarize the results in Section V.

II. Speaker Adaptation of HMM Parameters

In a speech recognition system based on phonetically-tied semicontinuous density HMM, the probability density function (pdf) of observing a feature vector o_t in state j is represented by

$$b_j(o_t)=\sum_{k=1}^{K} w_{jk}\, N(o_t\,;\mu_k,\Sigma_k). \qquad (1)$$

where w_{jk} is a weights for a Gaussian component, $N(o_t\,;\mu_k,\Sigma_k)$ is a Gaussian pdf with a mean vector μ_k and a covariance matrix Σ_k, and K is the number of codewords. The mean vectors and convariance matrices of the Gaussian densities used in summation of (1) constitute a codebook defined as

$$CB(j)\triangleq\{(\mu_1,\Sigma_1),(\mu_2,\Sigma_2),\cdots,(\mu_K,\Sigma_K)\}. \qquad (2)$$

The codebook is shared across among states belonging to the same position within each recognition subunit and having the same center phoneme. The Gaussian pdf's in the above equations are regarded as basic acoustic prototypes, namely senones [11]. We assumed that covariance matrices are diagonal. In this paper, we define a distribution as a weight vector for the codebook

$$w_j\triangleq(w_{j1},w_{j2},\cdots,w_{jK}). \qquad (3)$$

For an observation sequence $O=(o_1,o_2,\cdots,o_T)$ with length T, let $s=(s_1,s_2,\cdots,s_T)$ be the unobserved state sequence. The MAP estimation algorithm finds the parameter maximizing the posterior probability given observed samples or feature vectors as follows:

$$\hat{\Lambda}=\underset{\Lambda}{\operatorname{argmax}}\,[\max_{s} \Pr(\Lambda,s\,|\,O)] \qquad (4)$$

$$= \arg\max_{\Lambda} [\max_{s} \Pr(O, s|\Lambda) \Pr(\Lambda)] \quad (5)$$

where Λ denotes HMM parameters to be estimated, $\tilde{\Lambda}$ denotes the estimated HMM parameters, and $\Pr(\Lambda)$ is a prior density of Λ. Here, the HMM parameters Λ are defined as

$$\Lambda \triangleq (A, w, \mu, \Sigma) \quad (6)$$

where A denotes a state transition matrix of Markov chains. Assuming that the prior densities of the transition matrix, the distribution, and the codebook are independent each other, we can adapt the HMM prameters separately.

Without loss of generality, all the following formulations are derived assuming that feature vectors are one-dimensional and come from the same state of a Markov model. Assuming that the mean μ is random with prior density and the variance σ^2 is known and fixed, it is reported that the prior density for μ is also Gaussian with prior mean μ_0 and prior variance σ_0^2 [6]. A mean shift is defined as the difference between the MAP-estimated mean and the prior mean [5]. Then the mean shift in MAP estimation is calculated as [6], [2], [3], [5]

$$\Delta\mu \triangleq \tilde{\mu} - \mu_0 \quad (7)$$

$$= \frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2} (\bar{\mu}, \mu_0) \quad (8)$$

where n is the number of observed samples, $\tilde{\mu}$ is a an estimated mean, μ is a sample mean, and σ^2 is a sample variance. Here, we simplified calculation of the estimated mean shift as

$$\Delta\mu = \frac{n}{n+\alpha} (\bar{\mu} - \mu_0) \quad (9)$$

where α can be considered to be reflecting the ratio of the sample variance and the prior variance

$$\alpha = \frac{\sigma^2}{\sigma_0^2}. \quad (10)$$

By using α, there is no need to estimate hyperparameters of means.

We did not adapt the covariance matrices and the transition matrix because their contributions on system performance are not so significant considering computational complexity compared with means and distributions.

When prior densities for distributions are assumed as a form of the Dirichlet density, distributions are adapted as follows [3], [4]

$$\tilde{w}_{jk} = \frac{n_{jk} + \beta_{jk}}{\sum_{k=1}^{K} n_{jk} + \beta_{jk}} \quad (11)$$

where n_{jk} is the probabilistic count of observed samples assigned to codeword k in codebook j

$$n_{jk} = \sum_{t} \Pr(o_t \in \text{codeword } k \text{ of codebook } j | o_t, \Lambda) \quad (2)$$

$$= \sum_{t} \frac{w_{jk} N(o_t; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} w_{jk} N(o_t; \mu_k, \Sigma_k)} \Pr(o_t \in \text{state } j | \Lambda) \quad (13)$$

and β_{jk} is *a priori* estimate of distributions. Here, we calculated β_{jk} from the prior distribution value as

$$\beta_{jk} = \beta w_{jk}. \quad (14)$$

That is, we need not estimate the hyperparameters for distributions.

By using the simplified MAP-estimation algorithm for speaker adaptation, we have only to determine α and β instead of estimating all the hyperparameters of prior densities.

Ⅲ. A Vocabulary-Independent Speech Recognizer

We used 40 phoneme models including silence to build a vocabulary-independent isolated word recognizer. We clustered all Korean context-dependent subunits to 1,548 allophonic subunits based on an allophonic decision tree [12]. Each subunit except silence was modeled by a 3-state left-to-right HMM without skip transition. And the silence was modeled by a 1-state HMM. We used a total of 118 codebooks of size 50. A codebook was shared among states belonging to the same position of subunit models and having the same center phone.

To estimate initial HMM parameters of the vocabulary-independent speech recognizer, we used eight sets of 3,848 phonetically-optimized words(POW's) [13] pronounced by 64 speakers (32 males and 32 females). The counts of context-dependent subunits in the POW database was designed to follow the distribution of the counts in real speech. A set of POW's was divided into 8 partitions. Then, each partition was pronounced by a speaker and was recorded in a sound-proof booth.

We estimated HMM parameters of the vocabulary-independent speech recognizer in two steps. In the first step, context-independent models of the recognizer were initialized by using hand-labeled speech data and then trained by bootstrapping. Then, context-dependent subunit models were constructed from the estimated context-independent models and then trained also by bootstrapping. The recognizer obtained by these steps was used as a prototype recognizer in the following speaker adaptation experiments.

The speech signal was sampled at 16 kHz and segmented into 256-sample frames with each frame advancing every 160 samples. Each frame was parameterized by a 26-dimensional feature vector consisting of 13 perceptually linear prediction coefficients and their corresponding time derivatives. The recognition accuracy of the recognizer was 79.6% when two sets of POW's were used as test data.

Ⅳ. Experimental Results AND Discussion

A. Test and Adaptation Data

We used 75 phonetically-balanced words [4] pronounced by five male speakers as adaptation data in the supervised offline (batch) mode. For test data, we used 5 sets of 500 words pronounced by the same speakers. The test words consist of Korean railroad station names [14]. The speech data in both cases were recorded in a computer room, which is a different environment from the the sound-proof booth where the training data were recorded. The microphone used to record the test data was also different from the microphone used to record the training data.

When a speaker-independent recognizer was trained by using the 500 isolated word database pronounced by 36 speakers and tested by using the same kind of database pronounced by 12 speakers the speaker-independent speech recognizer showed the recognition accuracy of 83.6%.

B. Performance of the Vocabulary-Independent Speech Recognizer

As shown in Table 1, recognition accuracies of the vocabulary-independent and the vocabulary-dependent speech recognizers were 60% and 80%, respectively, when five speakers wever tested. That is, the number of errors in the vocabulary-independent speech recognizer increased by 50% compared with the vocabulary-dependent one. The large differences in recognition accuracies were caused by the fact that the recording environments and microphones used in the training and test data were different. The differences also come from the confusability of the test data. This is because because most of test words consist of only two syllables and they are often phonetically different from another wrods by one phoneme only. We note that interspeaker differences in recognition accuracies were also large. In our experiments, the speaker 'B' showed the lowest accuracy.

Table 1. Recognition Accuracies (%) of Vocabulary-Dependent and Vocabulary-Independent Speech Recognizers

Speech Recognizer	Speaker					Average
	A	B	C	D	E	
Vocabulary-Ind.	64	54	63	60	60	60
Vocabulary-Dep.	88	67	85	81	80	80

C. Speaker Adaptation Using the Same Vocabulary as Test Data

First, we adapted HMM parameters in a supervised offline mode using the same vocabulary as the test data.

Table 2. Recognition Accuracies (%) of Offline Adaptation Using the Same Vocabulary as the Test Data with Varying Number of Adaptation Words

No. Adapt. Words	codebook adapt.	distribution adapt.	codebook & distribution adapt.
10	61	60	59
20	64	61	64
40	67	63	67
80	74	65	76
160	81	69	82
320	81	70	81
500	89	76	92

To adapt codebooks and distributions of the speech recognizer, we adopted the segmental MAP algorithm [2] and used the values of $\alpha = 10.0$ and $\beta = 1.0$. We made no special efforts to optimize the values of α and β. Hence, other values may yield better recognition results. Table 2 shows recognition accuracies with varying the number of adaptation words when codebooks and/or distributions are adapted. To obtain the results, we used two sets of 500 test words pronounced by one speaker. Experimental results showed that codebook adaptation contributed to performance improvements more than distribution adaptation. When we performed both codebook and distribution adaptation with 80 adaptation wrods, the speech recognizer yielded recognition accuracy improved from 60% to 76%, or reduced 40% of recognition errors. With only 10 words used for adaptation, the recognition accuracy became worse. This means that the adaptation data were too short to estimate the HMM parameters reliably. When all words in the adaptation data set were used, the recognition accuracy was 92%. This can be regarded as the recognition accuracy of a speaker-dependent recognizer.

D. Speaker Adaptation Using Vocabulary Different from Test Data

Next, we performed supervised offline speaker adaptation using vocabulary different from test data. This experiment was to check whether the vocabulary-independent speech recognizer adapted to a new speaker by using a predefined adaptation word set can recognize different vocabulary in a different task with recognition accuracy comparable to the task-oriented speech recognizer. In this case, we can use a small-sized phonetically balanced word set as adaptation data. Table 3 shows recognition accuracies with varying number of adaptation words when codebooks and distributions are adapted, respectively. The experimental results showed that with this adaptation scheme, the performances became worse than the prototype recognizer. Combining codebook and distribution adaptation deteriorated the results form 60% to 43% when 75 phonetically-balanced words were used for adaptation. The performance deterioration was caused from the fact that HMM parameters trained by the training database were disturbed because the recognizer updated HMM parameters of all states sharing a codebook even though the subunits in the adaptation data belonging to the states have left-and right-contexts different from the subunits in the test data. That is, allophonic clustering and codebook sharing used in phonetically-tied semicontinuous HMM did harm to the speech recognizer with speaker adaptation.

Table 3. Recognition Accuracies (%) of Offline Adaptation of Codebooks and Distributions Using Vocabulary Different form the Test Data with Varying Number of Adaptation Words

No. Adapt.	Speaker					Average
Words	A	B	C	D	E	
10	60	47	59	56	54	55
20	56	38	53	51	50	50
30	56	38	52	50	46	48
40	54	37	49	47	44	46
50	53	38	48	46	42	45
60	52	34	48	45	41	44
75	49	35	43	43	43	43

E. Online Speaker Adaptation of Vocabulary-Independent Speech Recognizers

Finally, we performed experiments to analyze performance of a vocabulary-independent speech recognizer with unsupervised online speaker adaptation. Table 4 shwos recognition accuracies of online speaker adaptation with varying number of test words when codebooks and distributions are adapted in an online (sequential) mode. There was performance improvement from 60% to 77%, or 43% of error reduction compared with the prototype recognizer when 500 test words were used. We also observed the fact that improvement was mainly due to error reduction in speaker 'B' who showed the lowest recognition accuracy in the prototype recognizer.

Table 4. Recognition Accuracies (%) of Online Adaptation of Codebooks and Distributions with Varying Number of Test Words

No. Test	Speaker					Average
Words	A	B	C	D	E	
100	84	69	69	79	76	75
200	86	75	75	83	77	79
300	83	77	79	82	77	80
400	83	75	77	82	75	78
500	81	75	76	81	73	77

Table 5. Recognition Accuracies (%) of the Vocabulary-Dependent Speech Recognizer with Varying Number of Test Words

No. Test	Speaker					Average
Words	A	B	C	D	E	
100	91	77	86	86	89	86
200	90	76	88	86	87	85
300	90	74	89	88	88	86
400	89	70	86	85	82	82
500	88	67	85	81	80	80

Table 6. Recognition Accuracies (%) of Online Adaptation of Codebooks and Distributions for 100 Test

Test	Speaker					Average
Words	A	B	C	D	E	
1–100	84	69	69	79	76	75
101–200	84	80	80	87	78	82
201–300	91	81	88	81	78	84
301–400	71	70	72	79	69	72
401–500	75	75	70	77	64	72

As shown in Table 5, a task-oriented vocabulary-dependent speech recognizer yielded recognition accuracy of 80% when all of the 500 test words are used. This result shows that the task-oriented speech recognizer still yields slightly better recognition accuracy than the vocabulary-independent speech recognizer with speaker adaptation.

Tables 6 and 7 show recognition accuracies of the vocabulary-independent and the vocabulary-dependent speech recognizers for 100 test words with different time intervals when combined codebook and distribution

Table 7. Recognition Accuracies (%) of the Vocabulary-Dependent Speech Recognizer for 100 Test Words

Test Words	Speaker					Average
	A	B	C	D	E	
1-100	91	77	86	86	89	86
101-200	88	74	89	86	84	84
201-300	91	71	91	91	90	87
301-400	86	56	79	75	66	72
401-500	82	57	82	67	70	72

adaptation were performed, respectively. The recognition results in the tables can be regarded as instantaneous recognition accuracies while the results in Table 4 can be regarded as accumulated accuracies. The recognizer yielded lower accuracies in the last two intervals, which indicates that the intervals consisted of more confusing words. Tables 6 and 7 show that the vocabulary-independent speech recognizer achieved the recognition accuracy comparable to the vocabulary-dependent speech recognizer with sufficient adaptation[1] data.

V. Summary

We investigated performance of a vocabulary-independent speech recognizer with speaker adaptation. The vocabulary-independent speech recognizer used in this paper does not require a task-oriented speech database to estimate HMM parameters, but adapts the parameters recursively by using input speech data and the corresponding recognition results. We simplified formula of the MAP estimation algorithm to reduce the number of parameters to be estimated so that we need not estimate hyperparameters of the prior density. Experimental results showed that the vocabulary-independent speech recognizer with supervised offline speaker adaptation reduced 40% of recognition errors when 80 adaptation words from the same vocabulary as test data were used. But, the recognizer adapted by using vocabulary different from test data yielded lower recognition accuracy. This result was caused by allophonic clustering and codebook sharing used in phonetically-tied semicontinuous HMM. The speech recognizer with unsupervised online speaker adaptation reduced about 43% of recognition errors. And as recognition proceeded, its performance approached to that of a speaker-independent speech recognizer trained by a task-oriented speech database.

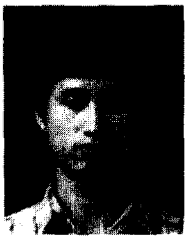
References

1. H.-R. Kim and H.-S.Lee, "Variable vocabulary word recognizer using phonetic knowledge-based model," *J. Acoust. Soc. Korea*, vol. 16, pp. 31-35, Feb. 1997.
2. C.-H.Lee, C.-H.Lin, and B.-H.Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806-814, Apr. 1991.
3. J.-L.Gauvain and C.-H.Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 291-298, Apr. 1994.
4. Q. Huo, C. Chan, and C.-H.Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 334-345, Sept. 1995.
5. G. Zavaliagkos, *Maximum a posteriori adaptation techniques for speech recognition*. PhD thesis, Northeastern Univ., Boston, MA, Oct. 1995.
6. M.H.DeGroot, *Optimal statistical decision*. New York: McGraw-Hill, 1970.
7. V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, pp. 294-300, July 1996.
8. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech, Language*, vol. 9, pp. 171-185, Apr. 1995.
9. K. Ohkura, H. Ohnishi, and M. Iida, "Speaker adaptation based on transfer vectors of multiple reference speakers," in *Proc. Int. Conf. Spoken Language Processing*, Tokyo, Japan, pp. 455-458, Sept. 1994.
10. B. M. Shahashahani, "A Markov random field approach to Bayesian speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, pp. 697-700, May 1996.
11. M.-Y. Hwang and X.Huang, "Subphonetic modeling with Markov states-senone," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Francisco, CA, pp. I-33-I-36, Mar. 1992.
12. M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, MN, pp. II-311-II-314, Apr. 1993.
13. Y.Lim and Y.Lee, "Implementation of the Pow (phonetically optimized words) algorithm for speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Deteroit, MI, pp. 89-92, May 1995.
14. J. R. Park, O. W. Kwon, D. Y. Kim, I. J. Choi, H. Y.

[1]The test data can be also viewed as adaptation data in the online adaptation mode.

Chung and C. K. Un, "Speech data collection for Korean speech recognition," *J. Acoust. Soc. Korea*, vol. 14, pp. 74-81, Aug. 1995.

▲Oh Wook Kwon



Oh Wook Kwon was born in Andong, Korea in 1964. He received the B.S. degree in electronic engineering from Seoul National University, in 1986, and the M.S. and Ph. D. degrees from Korea Advanced Institute of Science and Technology, in 1988 and 1997, respectively. Since 1988 he has been with Electronics and Telecommunications Research Institute, Taejon, Korea, as a senior member of technical staff in the Spoken Language Processing Section, working on spontaneous speech translation in multimedia environment and speech input/output processing for human-computer interface. His research interests include speech recognition, video coding, and adaptive signal processing.

▲Chong Kwan Un
Vol. 15, No. 4E, Dec. 1996.

▲Hoi-Rin Kim
Vol. 15, No. 4E, Dec. 1996.